I will explain from now on. The second model we used is Random Forest. I will explain the principle of this model briefly.

The concept you have to know before is bootstrap.

Bootstrap is creating data of the same size as the original data by Allowing duplication in given train data.

The order is as follows. We first create N bootstrap samples and use them to create N trees. Next the model trains each trees in ensemble. Then it selected the result of the prediction as a voting method.

10

The accuracy records 00point percent and the accuracy of actual data is oopointopercent

In the random forest Some data set wasn't be classified meaningfully, when using chosen features. So Random forest is not recommended if the number of features is large.

Random forest lost explanatory power which is advantage of decision tree, this is called a black boxed algorithm.

11.

The Third model is Gradient Boosting. What we spent most of our time in Gradient Boosting was to add new features from the given data. This is because GBM has more accuracy as we add more valid features. The features we added can be divided into four major categories. The first is to modify the incorrect data and specify the type of data. For example, if NA value is found in song length data, it is replaced with average value. and then we clarifies the data type of the feature where the existing type was object.

The second is Adding features related with song. we add score feature depending on the length of the song, Also we add the number of times the song was played.

The third is Adding features related with time. We subdivided the user's expiration time and registration time, and added the song release year using isrc which is the unique ID of the song.

The Fourth is Adding features related with song’s producer. We added features such as whether this song is featured and the number of artists, composers, lyricists and whether artists and composers are the same.

12

By using Gbm, the accuracy records eighty point six percents and the accuracy of actual test data was sixty eight point seven percent.

Important parameters of gbm are the learning rate and boosting frequency. when learning rate and boosting frequency were 0.2 and 100, The accuracy of the actual test data was highest. When we lower the the learning rate, the validation showed higher accuracy, but the actual test data showed overfitting with lower accuracy.

Also, gbm generally shows the accuracy increase when adding features. However, the accuracy of actual test data only increased when the feature was a valid.

13

here is conclusion. We tried to solve the problem by using these three models of machine learning. and the GBM was the best predictor among these models. In the case of Random Forest, the degree of overfitting was found to be lower than that of the other models. Neural networks were found to have poor prediction accuracy due to the nature of categorical data.

When we started project, we had no background on machine learning, but it was a good opportunity to deal with various data and models.

It was also impressive that we were able to analyze the large amount of customer data that the actual company had. I think this experience will be practical help for us.

14

Our presentation is over. Thank you for listening. Thank you.